

基于统计差分 LPP 的多模态间歇过程故障检测 *

郭金玉, 仲璐璐, 李 元

(沈阳化工大学 信息工程学院, 沈阳 110142)

摘 要: 针对工业过程数据存在的非高斯和多模态特性, 提出一种基于统计差分 LPP 的多模态间歇过程故障检测方法。首先将统计模量分析的方法应用到间歇过程训练数据集中, 计算统计过程变量的均值和方差, 将不等长的批次变成等长的统计量, 保证统计模量近似服从高斯分布; 然后运用差分算法使多模态变为单模态, 最后运用 LPP 算法进行降维和特征提取, 计算样本的 T^2 统计量, 并利用核密度估计确定控制限。对于新来的测试样本数据经统计差分处理后, 向 LPP 模型上进行投影, 计算新数据的 T^2 统计量并与控制限比较进行故障检测。最后通过半导体过程数据的仿真结果表明, 该算法的故障检测效果最好, 验证了所提方法的有效性。

关键词: 多模态间歇过程; 统计模量分析; 差分算法; 局部保持投影算法; 故障检测

中图分类号: TP277 **doi:** 10.3969/j.issn.1001-3695.2017.07.0665

Fault detection of multi-mode batch process based on statistics difference LPP

Guo Jinyu, Zhong Lulu, Li Yuan

(College of Information Engineering, Shenyang University of Chemical Technology, Shenyang 110142, China)

Abstract: Aiming at non-Gaussian and multi-mode characteristics exist in industrial process data, this paper proposed a fault detection of multi-model batch process method based on statistics difference LPP. Firstly, the method of statistical pattern analysis was applied to the batch process training data set to calculate the mean and variance of statistical process variables, and turned the uneven-length batches into equal-length statistics. It could ensure that the statistics pattern approximately obeyed the Gaussian distribution. Then it used the difference algorithm to transform the multi-mode into single mode. Finally, it used the LPP algorithm to reduce dimension and extract feature, and calculated the T^2 statistic of the sample. And it used the kernel density estimation to determine the control limit. The new test sample data projected onto the LPP model after statistics difference processing. Then it calculated the T^2 statistics of the new data and compared them with the control limit for fault detection. Finally, the simulation results of the semiconductor process data show that this algorithm has the best fault detection effect, and demonstrate the effectiveness of the proposed algorithm.

Key Words: multimode batch process; statistics pattern analysis; difference algorithm; locality preserving projections algorithm; fault detection

0 引言

随着现代工业的快速发展, 原材料的波动、工作点的调整、产品规格及批次不同等使生产过程的工况频繁发生改变, 这就导致了过程变量不完全服从高斯分布, 且其均值与协方差结构将随着模态的切换而发生变化。由于传统主成分分析法^[1~3](principal component analysis, PCA)应用于多模态过程中, 其需要数据满足单一分布的基本特点无法满足, 同时也缺乏提供非高斯数据特征的能力, 所以如果直接将 PCA 算法应用到间歇过程中并不能得到令人满意的检测结果。

针对工业生产过程中存在的多模态问题, Zhao 等人^[4~5]提出了基于 PCA 和 PLS 的多模型方法, 该方法通过对过程中的多个模态分别建立局部模型进行故障检测, 但该方法具有一定的局限性, 当建立多个子模型之后, 需要对实时样本实现在线检测, 这时对在线样本的归属划分及其重要, 错误的判别会导致模型与检测样本不匹配, 产生错误的检测结果。为此, 华东理工大学的马贺贺等人^[6]提出了一种局部近邻标准化数据处理方法(local neighborhood standardization, LNS), 利用每个样本局部近邻的均值和标准差标准化该样本, 使数据标准化后尽可能服从单峰分布, 并结合 PCA 算法提出了一种多模态故障检测

基金项目: 国家自然科学基金重大项目(61490701); 国家自然科学基金资助项目(61174119, 61673279); 辽宁省教育厅重点实验室项目(LZ2015059); 辽宁省自然科学基金资助项目(201602584); 辽宁省教育厅资助项目(L2016007, L2015432)

作者简介: 郭金玉(1975-), 女, 山东高唐人, 副教授, 博士, 主要研究方向为故障诊断、生物识别算法及应用(shandong401@sina.com); 仲璐璐(1992-), 女, 硕士研究生, 主要研究方向为故障诊断; 李元(1964-), 女, 教授, 博士, 主要研究方向为基于数据驱动技术的复杂过程故障检测与诊断。

方法: 局部近邻标准化主元分析算法 (local neighborhood standardization principle component analysis, LNS-PCA)。该算法是将过程中所有模态视为一个统一整体, 建立统一的故障检测模型, 可以有效避免对实时检测样本归属的判别, 但该方法的近邻数 k 的选取对检测结果影响较大, 而且该方法根据经验选取的近邻数 k 需要以各个模态过程知识为依托。为了弥补 LNS-PCA 算法的不足, Wang 等人^[7]提出了一种加权 k 近邻标准化 PCA (weighted K neighbourhood standardisation PCA, WKNS-PCA) 算法, 该方法可以在无须确定近邻参数 k 的情况下使多模态过程数据近似服从高斯分布, 和 LNS-PCA 算法相比, 该算法更加科学有效, 具有更优越的数据处理能力, 但以上两种算法都是全局算法, 不能保持数据的局部结构。为了保持数据的局部结构, Hu 等人^[8]将局部保持投影 (locality preserving projections, LPP) 算法成功应用在统计过程监控中。为了更好地保持数据的局部结构, Cai 等人^[9]提出了正交局部保持投影 (orthogonal locality preserving projections, OLPP), 在 LPP 的基础上增加了一个正交化的约束条件, 通过迭代计算得到相互正交的投影方向。在此基础上, Guo 等人^[10]提出了一种基于动态多向正交局部保持投影 (dynamic multiway orthogonal locality preserving projections, DMOLPP) 算法用于间歇过程故障检测, 该算法将滑动窗口技术与正交局部保持投影相结合, 能够在保留原始训练样本特征信息的同时降低数据误差重构方面的难度。然而这些局部算法不能处理数据的多模态问题。

针对数据信息不全面以及多工况所引起的非高斯、多模态等问题, 本文尝试将多模态问题转换为单模态问题进行故障检测。为了使数据从多模态变为单模态, 并且近似服从高斯分布, 在提高不等长间歇过程故障检测性能的同时, 降低算法的复杂度, 本文结合统计模量分析 (statistics pattern analysis, SPA) 算法, 提出一种基于统计差分 LPP (statistics difference locality preserving projections, SDLPP) 的多模态间歇过程故障检测方法。其思想是运用 SPA 算法保证数据近似服从高斯分布, 然后利用差分算法, 在保持数据内部结构的同时, 使多模态数据变为单模态, 最后运用 LPP 算法提取数据特征并保持数据局部结构, 从而达到提高多模态间歇过程故障检测性能的目的。

1 局部保持投影算法

LPP 算法^[11-14]是一种用于提取数据特征信息的降维方法。它可以很好的保留数据的局部信息, 主要考虑的是保持数据中近邻点之间的结构, 其本质是对拉普拉斯特征映射的线性逼近。它是 LE 算法的扩展, 且与 LE 的目标函数相同, 但是它使用显式的线性映射。算法的核心是寻找转换矩阵 A 使一系列的 $X = [x_1 \ x_2 \ \dots \ x_n] \in R^{m \times n}$ 投影到 $Y = [y_1 \ y_2 \ \dots \ y_n] \in R^{l \times n} (l = m)$ 上, 即 $y_i = A^T x_i$, 使 Y 尽可能代表 X 。求解 A 可以通过优化如下最小值问题:

$$\begin{aligned} a_{opt} &= \arg \min_a \sum_{i,j} \|y_i - y_j\|^2 W_{ij} \\ &= \arg \min_a \sum_{i,j} \|a^T x_i - a^T x_j\|^2 W_{ij} \\ &= \arg \min_a a^T X L X^T a \end{aligned} \quad (1)$$

约束条件为

$$a^T X L X^T a = 1 \quad (2)$$

其中: L 是 Laplacian 矩阵, $L = D - W$, W 是定义在数据点上的相似矩阵, $D_{ii} = \sum_j W_{ij}$, 其计算方式如下:

$$W_{ij} = \begin{cases} \exp(-\|x_i - x_j\|/t), & \text{如果 } x_i \text{ 是 } x_j \text{ 的 } p \text{ 近邻} \\ & \text{或 } x_j \text{ 是 } x_i \text{ 的 } p \text{ 近邻} \\ 0, & \text{其它} \end{cases} \quad (3)$$

其中: W_{ij} 是加权矩阵, t 的设定根据经验选择。求目标函数的最小值可以保证近邻点 x_i 和 x_j 的投影 y_i 和 y_j 也是近邻点。寻求最佳投影矩阵 A 转换为求广义特征问题的最小特征值所对应的特征向量:

$$X L X^T A = \lambda X D X^T A \quad (4)$$

其中: $X L X^T$ 和 $X D X^T$ 都是对称且半正定的, 因此求矩阵 $(X D X^T)^{-1} X L X^T$ 的最小特征值对应的特征向量, 即得到投影矩阵 A 。

2 基于统计差分 LPP 的多模态间歇过程故障检测

2.1 统计模量分析

基于统计模量分析^[15-18]的方法是用统计特征值组成的统计模量矩阵来重新定义和衡量原始间歇过程数据, 本文所用到的统计特征包含均值和方差。假定每个批次的均值为 μ , 方差为 ν , 一个批次所有的统计特征组合成一个 $(1 \times 2m)$ 维的特征向量 $SP = [\mu_1 \nu_1 \mu_2 \nu_2 \ \dots \ \mu_{m-1} \nu_{m-1} \mu_m \nu_m]$ 。由此可见, 最终由所选定的特征值叠加而成的行矢量 SP 是等长的。而用于建模的统计模量 SPs 则是由 k 个批次叠加在一起所构成的训练矩阵。统计模量 SPs 能够提取间歇过程特征, 包括过程的非线性和非高斯性。

2.2 差分算法

差分算法能够剔除数据的多模态结构。对数据矩阵 $X(m \times n)$ (m 代表采样次数, n 代表测量变量个数) 中第 i 个样本 $X_{i,n}$, 找到该样本的最近邻 $X_{k,n}$, 然后进行差分运算, 计算方式如下:

$$DX = X_{i,n} - X_{k,n} \quad (5)$$

其中: $DX(m \times n)$ 是差分矩阵。

通过一个人工合成的多模态数值例子比较差分之前和差分之后的结果, 验证差分算法可以剔除数据多模态的有效性。数值例子中每个样本有两个变量, 第一个变量 x_1 服从 $[0,1]$ 均匀分布, 第二个变量 x_2 则与变量 x_1 线性相关, 具体数据来源于式 (6)。

$$x_2 = x_1 + \text{noise} \quad (6)$$

通过适当的位置转换, 获得 2 个模态的数值例子。图 1 是

两个模态的 400 个样本所构成的数据分布散点图。由此可见, 所用到的数值例子是多模态的。对该数值例子进行差分运算后得到的数据分布散点图如图 2 所示。通过图 1 和 2 的比较可以很直观地看出, 差分算法可以使数据的多模态结构转换为单模态结构, 从而验证了该算法的有效性。

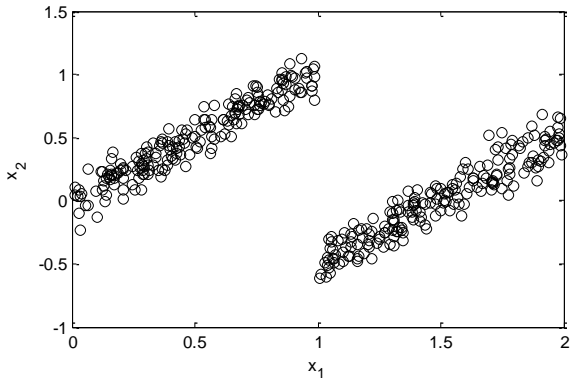


图 1 数据分布散点图

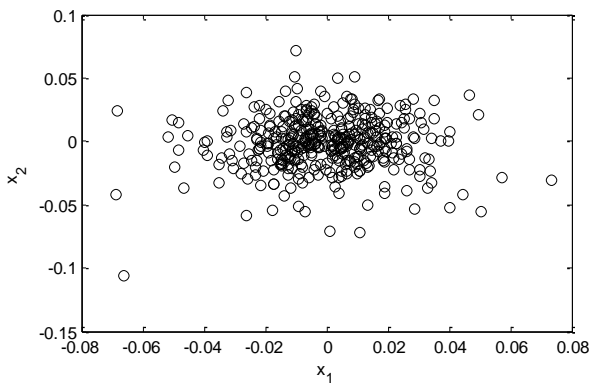


图 2 差分后的数据分布散点图

2.3 基于统计差分 LPP 的多模态间歇过程故障检测

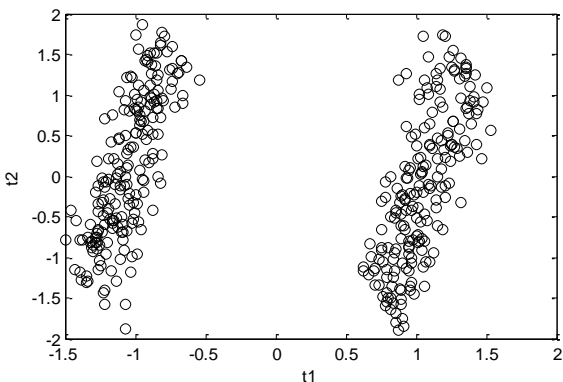


图 3 LPP 算法处理后的数值例子主元图

用 LPP 算法对上述所提到的数值例子进行处理, 得到如图 3 所示的主元图。由图 3 的主元图可以看出, 经过 LPP 算法处理后仍然是 2 个模态的, 这就说明了 LPP 算法无法剔除多模态结构。由于差分算法可以剔除多模态结构, 所以将统计模量分

析、差分算法和局部保持投影算法相结合, 提出了基于统计差分 LPP 的多模态间歇过程故障检测算法。在保证数据在近似服从高斯分布的条件下, 利用差分算法剔除多模态结构, 为 LPP 提供所需要的数据特点。

基于统计差分 LPP 的多模态间歇过程故障检测包括建立模型和故障检测两个部分, 该算法的整体流程图如图 4 所示。具体步骤如下:

a) 建立正常状态下的模型

(a) 收集正常操作时的间歇过程历史数据集 $\mathbf{X}_k (I \times J \times d)$, 其中 I 是批次数, J 是变量数, d 是一个批次过程总的反映时刻数。

(b) 计算每个批次 $\mathbf{X}_k (I \times J \times d)$ 的均值和方差, 得到行矢量 \mathbf{SP} ;

(c) 将 I 个批次数据得到的行矢量 \mathbf{SP} 叠加到一起, 得到统计模量 \mathbf{SPs} ;

(d) 对统计模量 \mathbf{SPs} 的每个样本, 找到其最近邻, 将该样本与其最近邻进行差分运算得到统计差分矩阵 \mathbf{SD} ;

(e) 选择合适的参数, 运用 LPP 算法获取投影矩阵 \mathbf{A} , 并计算统计量 T^2 , 建立正常工况下的 LPP 模型。

(f) 利用核密度估计的方法确定统计量 T^2 的 95% 控制限。

b) 故障检测

(a) 新来一个时刻 k 的批次数据 $\mathbf{X}_k^{new} (J \times d)$, 计算其均值和方差, 得到 \mathbf{SP}^{new} ;

(b) 在建模数据中寻找 \mathbf{SP}^{new} 的最近邻样本, 将 \mathbf{SP}^{new} 与其最近邻样本进行差分运算, 得到统计差分数据矩阵 \mathbf{SD}^{new} ;

(c) 新样本的差分矩阵向 LPP 模型上进行投影, 并计算统计量 T^2 ;

(d) 通过比较 \mathbf{X}_k^{new} 的 T^2 是否超过控制限来判断当前过程是否正常, 如果 \mathbf{X}_k^{new} 超出控制限则为故障样本, 否则 \mathbf{X}_k^{new} 为正常样本。

3 仿真实例

为了验证本文所提出算法的有效性, 分别将 PCA、LPP、DPCA、DLPP、SPCA、SLPP、SDPCA 和 SDLPP 八种算法分别应用于半导体生产过程数据的故障检测中, 并进行了对比, 进一步阐明 SDLPP 算法在对多模态间歇过程数据进行故障检测时所具有的优越性。

半导体生产过程作为一个完善的工业过程仿真平台, 在基于数据驱动的故障检测研究领域得到了广泛的应用。本文应用半导体工业实例—A1 堆腐蚀过程^[19-22]比较不同的故障检测方法的性能。半导体生产过程是典型的非线性、时变、多阶段和多模态的间歇过程。半导体生产过程数据由 108 个正常硅片和 21 个故障硅片构成。由于两个批次过程 (第 56 个正常批次和第 12 个故障批次) 丢失大量的数据, 所以实际的批次为 107 批正常数据和 20 批故障数据。在该仿真实验中, 随机抽取 96 个正常批次为建模数据, 其余 11 个正常批次为校验数据, 故障批

次为20个。从 21 个测量变量中选取 17 个变量作为检测变量, 如表 1 所示。每个批次是不等长的, 持续时间在 95-112 秒之间变化。

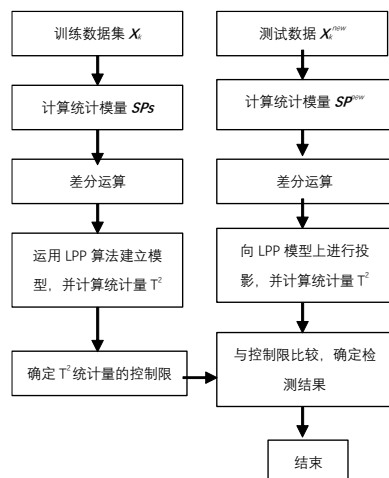


图 4 基于统计差分 LPP 的多模态间歇过程故障检测流程图

表 1 半导体生产过程所用的检测变量

序号	过程变量	序号	过程变量
1	BC13 流量	10	RF 功率
2	C12 流量	11	RF 阻抗
3	RF 底部功率	12	TCP 调谐
4	A 检测端点	13	TCP 相位误差
5	氢压力	14	TCP 阻抗
6	Helium 压强	15	TCP 顶部功率
7	RF 调谐	16	TCP 负荷
8	RF 负荷	17	Vat 阀门
9	相对误差		

正如前面所指出的, 半导体生产过程是典型的非线性、时变、多阶段和多模态的间歇过程。1 个例子 (变量 EndPtA) 表明了这些特性, 如图 5 所示。

为了验证了 SPA 算法可以解决生产过程中存在的非高斯问题, 将 SPA 算法应用到半导体不等长多模态间歇过程中所有批次的第 2 个变量中, 如图 6 所示。由图 6 可以清楚地看出, 原始变量并不服从高斯分布, 但经过 SPA 处理过的原始变量的均值

和方差都近似服从高斯分布, 这就验证了该算法的有效性。

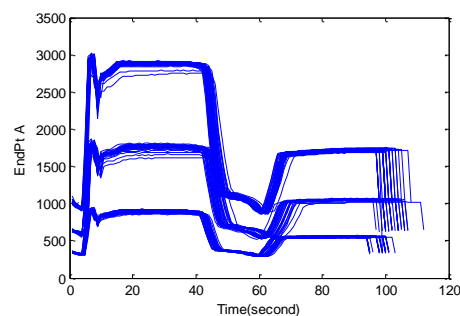
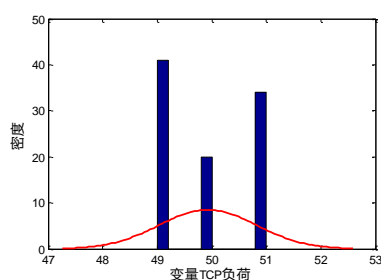
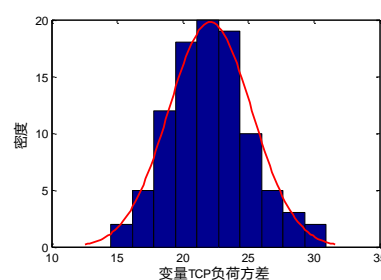
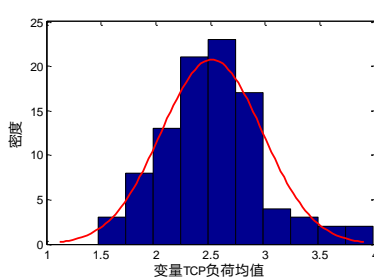


图 5 变量 EndPt A

本文对 96 个不等长的正常批次数据分别运用 PCA、LPP、DPCA、DLPP、SPCA、SLPP、SDPCA 和 SDLPP 方法进行建模, 并对 11 个校验批次和 20 个故障批次数据进行故障检测。八种算法的主元个数均为 16。八种方法的 T^2 检测结果如图 7 所示。图中的虚线是 T^2 统计量的 95% 控制限。由图 7 可以看出在主元个数选取条件相同的情况下, 运用 PCA 算法进行故障检测时, 校验数据全部检测出来, 故障数据有 15 个没有检测出来。运用 LPP 算法时, 校验数据全部检测出来, 故障数据有 15 个没有检测出来。运用 DPCA 算法进行检测时, 校验数据全部检测出来, 故障数据有 13 个没有检测出来。运用 DLPP 算法进行检测时, 校验数据全部检测出来, 故障数据有 16 个没有检测出来。运用 SPCA 算法时, 校验数据有 3 个没有检测出来, 而故障数据有 13 个没有检测出来。运用 SLPP 算法进行检测时, 校验数据有 3 个没有检测出来, 而故障数据有 5 个没有检测出来。运用 SDPCA 算法时, 校验数据全部检测出来, 而故障数据有 6 个没有检测出来。运用 SDLPP 算法进行检测时, 校验数据有 1 个没有检测出来, 故障数据也有 1 个没有检测出来。与传统 LPP 算法相比, SLPP 算法的检测效果要明显优于传统 LPP, 验证了统计模型量分析算法的有效性。与 SPCA 和 SDPCA 算法相比, SLPP 和 SDLPP 算法的故障检测性能分别要优于 SPCA 和 SDPCA 算法。这是由于 SPCA 和 SDPCA 算法提取的是数据的全局特征, 而 SLPP 和 SDLPP 算法提取的是数据的局部特征, 提高了故障检测性能。与其余七种算法相比, SDLPP 算法的检测效果最佳, 验证了 SDLPP 算法在多模态间歇过程故障检测中的有效性。



(a) 原始变量分布直方图



(b) SPA 后的变量分布直方图

图 6 变量分布直方图

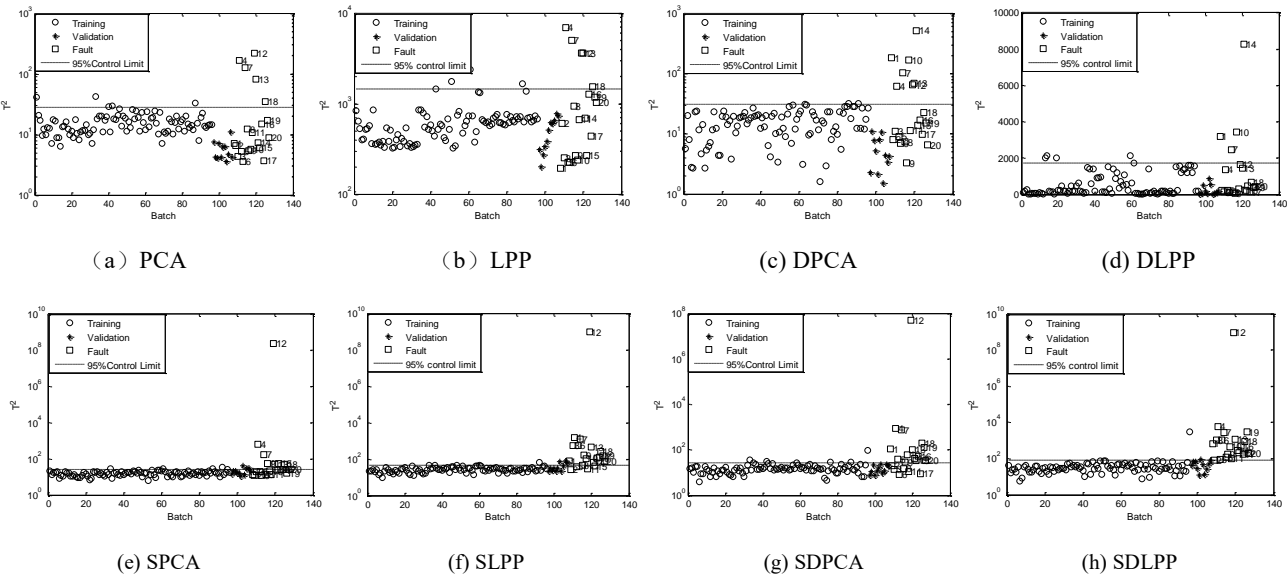


图 7 八种算法对半导体过程数据的检测结果

表 2 是八种算法对半导体过程数据的具体检测结果统计。由表 2 可以看出, 对多模态过程进行故障检测, SDLPP 算法能保证在最低的漏报率下, 误报率相对较低。与其他七种算法相比, SDLPP 的故障检测效果最好, 验证了该算法的有效性。

表 2 半导体过程数据检测结果统计

算法	误报率/%	漏报率/%
PCA	0	75
LPP	0	75
DPCA	0	65
DLPP	0	80
SPCA	27	65
SLPP	27	25
SDPCA	0	30
SDLPP	9	5

4 结束语

本文提出一种基于统计差分 LPP 的多模态间歇过程故障检测方法, 并应用于半导体生产过程中进行过程监控和故障检测。该方法首先将多模态间歇过程的统计量作为训练集, 然后对训练集进行差分处理, 对其进行多模态特征分析, 最后应用 LPP 方法对差分后的训练集进行故障检测, 有效地解决了数据信息不全面, 以及多工况引起的非高斯、多模态的问题。半导体生产过程是一个典型的多模态间歇过程, 通过与 PCA、LPP、DPCA、DLPP、SPCA、SLPP 和 SDPCA 等方法的仿真研究进行对比, 验证了该方法对多模态间歇过程进行故障检测的优越性。

参考文献:

[1] 周东华, 李钢, 李元. 数据驱动的工业过程故障检测与诊断技术 [M]. 北京: 科学出版社, 2011: 1-76.

[2] Wold S, Esbensen K, Geladi P. Principal component analysis [J]. Chemometrics and Intelligent Laboratory Systems, 1987, 2 (1/2/3): 37-52.

[3] 许仙珍, 谢磊, 王树青. 基于 PCA 混合模型的多工况过程监控 [J]. 化工学报, 2011, 62 (3): 743-752.

[4] Zhao Shijian, Zhang Jie, Xu Yongmao. Performance monitoring of processes with multiple modes through multiple PLS models [J]. Journal of Process Control, 2006, 16 (7): 763-772

[5] Zhao Shijian, Zhang Jie, Xu Yongmao. Monitoring of processes with multiple operation modes through multiple principle component analysis models [J]. Industrial & Engineering Chemistry Research, 2004, 43 (22): 7025-7035.

[6] Ma Hehe, Hu Yi, Shi Hongbo. A novel local neighborhood standardization strategy and its application in fault detection of multimode processes [J]. Chemometrics and Intelligent Laboratory Systems, 2012, 118(7): 287-300.

[7] Wang Guozhu, Liu Jianchang, Zhang Yingwei, et al. A novel multi-mode data processing method and its application in industrial process monitoring [J]. Journal of Chemometrics. 2015, 29 (2): 126-138.

[8] Hu Kunlun, Yuan Jingqi. Multivariate statistical process control based on multiway locality preserving projections [J]. Journal of Process Control, 2008, 18 (7): 797-807.

[9] Cai Deng, He Xiaofei, Han Jiawei, et al. Orthogonal laplacianfaces for face recognition [J]. IEEE Trans on Image Processing, 2006, 15 (11): 3608-3614.

[10] Gou Jinyu, Qi Leilei, Li Yuan. Fault detection of batch process using dynamic multi-way orthogonal locality preserving projections [J]. Journal of Computational Information Systems, 2015, 11 (2): 577-586.

[11] He Xiaofei, Yan Shuicheng, Hu Yuxiao, et al. Face recognition using Laplacian faces [J]. IEEE Trans on Pattern Analysis and Machine Intelligence, 2005, 27 (3): 328-340.

[12] Belk M, Niyogi P. Laplacian eigenmaps and spectral techniques for

- embedding and clustering [C]// Proc of the 14th International Conference on Neural Information Processing Systems: Natural and Synthesi. Cambridge: MIT Press, 2001: 585-591.
- [13] Fadi D, Ammar A. Enhanced and parameterless locality preserving projections for face recognition [J]. Neurocomputing, 2013, 99 (1): 448-457.
- [14] Zheng Heng, Liu Jijian, Wu Chaoxia, et al. A new construction method of neighbor graph for locality preserving projections [J]. Journal of Information and Computational Science (S1548-7741), 2013, 10 (5): 1357-1365.
- [15] 郭金玉, 赵璐璐, 李元. 基于统计特征的不等长间歇过程故障诊断研究 [J]. 计算机应用研究, 2014, 31 (1): 128-130.
- [16] 逢玉俊, 李娜, 李元, 等. 基于统计模式分析的多变量连续过程故障检测 [J]. 计算机应用研究, 2015, 32 (7): 2060-2064.
- [17] 张成, 李元. 基于统计模量分析间歇过程故障检测方法研究 [J]. 仪器仪表学报, 2013, 34 (9): 2103-2110.
- [18] He Fei, Xu Jinwu. A novel process monitoring and fault detection approach based on statistics locality preserving projections [J]. Journal of Process Control, 2016, 37 (5): 46-57.
- [19] Wise B M, Gallagher N B, Butler S W, et al. A comparison of principal component analysis, multiway principal component analysis, trilinear decomposition and parallel factor analysis for fault detection in a semiconductor etch process [J]. Chemometrics, 1999, 13 (3-4): 379-396.
- [20] Lee S P, Chao A K, Tsung F, et al. Monitoring batch processes with multiple on-off steps in semiconductor manufacturing [J]. Journal of Quality Technology, 2011, 43 (2): 142-157.
- [21] He Q P, Wang Jin. Fault detection using the k-nearest neighbor rule for semiconductor manufacturing processes [J]. IEEE Trans on Semiconductor Manufacturing, 2007, 20 (4): 345-354.
- [22] Yu Jianbo. Fault detection using principal components based Gaussian mixture model for semiconductor manufacturing processes [J]. Semiconductor Manufacturing, 2011, 24 (3): 432-444.